TextCLIP: Text-Guided Face Image Generation And Manipulation Without Adversarial Training

Anonymous author(s)
Submission Id: xxxx

Figure 1: In this work, we propose TextCLIP, a unified framework for text-guided image generation and manipulation without adversarial training. On the left is the comparison between TextCLIP and baseline [42] on image generation task, on the right are the results of TextCLIP on image manipulation task.

ABSTRACT

Text-guided image generation refers to the generation of a corresponding image based on a specified text, while text-guided image manipulation refers to semantically edit parts of a given image based on a specified text. For these two similar tasks, the key point is to ensure image fidelity as well as semantic consistency. Many current approaches require complex multi-stage generation and adversarial training, while struggling to provide a unified framework for both tasks. In this work, we propose TextCLIP, a unified framework for text-guided image generation and manipulation without adversarial training. The proposed method accepts input from images or random noise corresponding to these two different tasks, and under the condition of a specific text, a carefully designed mapping network that exploits the powerful generative capabilities of StyleGAN and the text image representation capabilities of Contrastive Language-Image Pre-training (CLIP) generates images of up to 1024×1024 resolution that can currently be generated. Extensive experiments on the Multi-modal CelebA-HQ dataset have demonstrated that our proposed method outperforms existing state-of-the-art methods, both on text-guided generation tasks and manipulation tasks.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Text-guided image generation, Text-guided image manipulation, StyleGAN

1 INTRODUCTION

Text-guided image generation and manipulation has recently gained significant attention and made some progress in the field of computer vision [8, 26, 29, 42, 46]. Text-guided image generation and manipulation require the generation or modification of images based on specified text, which is two complex cross-modal tasks. Text and images belong to two different modalities, and cross-modal data operation is difficult. For the task of text-guided image generation, Reed *et al.* [30] first proposed text-guided image generation using adversarial generative networks [11] and generated more research on text-guided image generation [12, 23, 40, 45–48, 52, 54]. The generated image needs to not only produce a sufficiently realistic image, but also be semantically consistent with the corresponding text. Some previous research has focused on multi-stage generation, where multiple low-quality images are first generated to produce high-quality images, which means that multiple generators and discriminators need to work together. These efforts require a tedious multi-stage generation process and complex adversarial training, which is very time-consuming and difficult to train. For other recent works [8, 29, 42, 46], some of them have much room for improvement in the quality of the generated images, while the others require a large number of training parameters or training data, making training too expensive.

Table 1: Comparison of Different Text-Guided Image Generation Models.

Method	AttnGAN [45]	ControlGAN [23]	DAE-GAN [32]	XMC-GAN [46]	TediGAN [42]	**TextCLIP**
One Generator	-	-	-	✓	✓	✓
Single Model	✓	✓	✓	✓	-	✓
High Resolution	-	-	-	-	✓	✓
Manipulation	-	✓	-	-	✓	✓
Open World	-	-	-	-	✓	✓
w/o Adversarial Training	-	-	-	-	✓	✓

For text-guided image manipulation task [5, 14, 17, 35, 51], the corresponding image needs to be modified according to the specified text. It is important to note that the areas of the image that are semantically irrelevant to the specified text should be kept as close as possible to the original image, and only those areas of the image that are semantically relevant should be modified.TediGAN [42] is the first work to provide text-guided image generation and manipulation by exploiting the semantic properties of the latent space of GAN. However, the performance of TediGAN has much room for improvement.

StyleGAN [19–22] is now state-of-the-art generative adversarial networks with powerful image generation capabilities, providing realistic images at resolutions up to 1024×1024, and more importantly, StyleGAN's latent space with good semantic performance and unraveling capabilities. The latent space of StyleGAN has been the subject of much recent research progress [1, 34, 41], which has significantly advanced several fields. Contrastive Language-Image Pretraining (CLIP) [28] is a powerful multimodal pretrained model that provides powerful text image representation capabilities and can be used as a supervisor for cross-modal tasks to achieve semantic-visual alignment. Some meaningful works based on pretrained StyleGAN and CLIP have been born recently [2, 3, 10, 25, 26, 33, 43, 50].

In this work, we propose TexCLIP, a unified framework for text-guided image generation and manipulation without adversarial training, which doesn't require the complex multi-stage generation and tedious adversarial training and outperforms extant state-of-the-art methods in two tasks. First, either random noise or images are used as input, with the random noise corresponding to the text-guided image generation task and the images corresponding to the text-guided image manipulation. Using a pre-trained encoder, the input is transformed into w_0, which is used as the initial latent code. w_0 is then subjected to a level-channel mapper with two parts: (a) level mapper: from coarse to fine, divided into three separate networks (coarse, medium, fine), each mapping a part of the initial latent code w_0. (b) channel mapper: consists of 18 style modulation networks. The final mapping latent code w_t is obtained by level-channel mapper, which is then processed differently with the initial latent code w_0 for different tasks to obtain the final latent code w_s. w_s is used as input to the generator of StyleGAN to obtain the final image. Table 1 shows how our method compares with other methods. Compared with other text-guided image generation methods, our proposed method is able to produce high-resolution images, support manipulation of images and accept open-world text as input without the need for adversarial training and multi-stage generation. In contrast to TediGAN [42], we do not need to train different models for different texts.

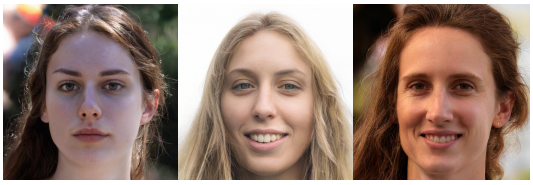
In summary, this work consists of the following main contributions:

- For the two distinct tasks of text-guided image generation and text-guided image manipulation, we propose TextCLIP, a unified framework that enables text-guided image generation and manipulation without the need for complex adversarial training.
- We propose level-channel mapper that uses text as a condition to semantically map the initial latent code to the latent space $\mathcal{W}+$ [1] of StyleGAN. Compared to previous work TediGAN [42], level-channel mapper does not require training different networks for different text conditions.
- Extensive qualitative and quantitative studies have shown that our proposed TextCLIP outperforms existing state-of-the-art methods on these two different tasks.

2 RELATED WORK

2.1 Text-Guided Image Generation

We have divided previous work on text-guided image generation into two categories. The first category is multi-stage generative models, where multiple generators and discriminators need to be used to complete the text-guided image generation work. StackGAN [47] was the first multi-stage generative model that used multiple generators and discriminators to first generate low-quality images and then generate high-quality images. Later StackGAN++ [48] implemented end-to-end training based on StackGAN to generate higher quality images. AttnGAN [45] introduced an attention mechanism to achieve word-level image generation, generating more realistic and realistic high-quality images; in addition, the proposed Deep Attention Multimodal Similarity Model (DAMSM) to compute the similarity of image-text pairs. DM-GAN [54] ugenerates a low-resolution initial image with a smaller model size, and then uses dynamic memory networks to purify the initial image to produce a more realistic image. Much subsequent work, optimised on the basis of AttnGAN, has achieved higher quality image generation with more accurate semantic alignment [4, 23, 27, 32].ControlGAN [23] proposes an innovative multi-stage generation architecture and introduces perceptual loss to solve the problem that if some words in a sentence are changed during text-guided image generation, the composite image will be very different from the original image.MirrorGAN [27] is inspired by CycleGAN [53] and reduces the generated images to text, further improving the quality of the generated images. DAE- GAN [32] takes into account the 'aspect' information of the input text and incorporates it into the multi-stage generation process.



The woman has long hair and double eyelids

The man has white skin and no beard

Figure 2: Diverse text-guided image generation results. On the same text conditions, TextCLIP can generate multiple images at 1024×1024 resolution.

The second category is represented by XMC-GAN [46] and DALL-E [29]. XMC-GAN [46] uses contrast learning as supervision, takes into account image text contrast loss, true-false image contrast loss, and image region word contrast loss, and uses modulation layers to build a single-stage generative network that achieves state-of-the-art performance on several public dataset. DALL-E [29] trains a large number of text-image pairs on a Transformer with 12 billion network parameters, achieving zero-s hot generation. CogView [8] is similar to DALL-E in that it trains a Transformer [38] with 4 billion network parameters to autoregressively model images and text, achieving stronger zero-sample generation. TediGAN-A [44] uses a pretrained StyleGAN with a GAN inverse module, a visual semantic similarity module, and an instance-level optimization module to perform an optimized search in the latent space, resulting in text-guided image generation. TediGAN-B [43] improves the performance of TediGAN-A by using a pretrained image inversion model and CLIP [28].

2.2 Text-Guided Image Manipulation

ManiGAN [24] is a multi-stage text-guided image manipulation work using multiple generators and discriminators and has demonstrated good performance on the CUB and COCO datasets. StyleCLIP [26] provides three different methods for text-guided image manipulation, including optimizers, mappers and global direction. The mapper requires training a model with different parameters for different text conditions and is an inflexible approach for practical applications. The optimizer and global direction approaches require inference on different instances each time and take longer to infer. Our proposed TextCLIP differs from previous work in that we propose a more flexible way to perform text-guided image manipulation and achieve higher quality image manipulation. Instead of training a different model for each text, TextCLIP can use the trained model to generate results directly based on the image and text conditions, without excessive inference time.For example, we can train a model for the same class of text conditions, e.g. a model trained about skin color can perform inference on dark skin, white skin, red skin, etc.

2.3 StyleGAN And CLIP

StyleGAN [19–22] is an excellent tool for image generation and is state-of-the-art work in the field of adversarial generative networks. StyleGAN's input is mapped to the latent space by processing eight fully connected layers, which are then fed into the StyleGAN generator. The StyleGAN generator has 18 layers, with every two layers corresponding to a resolution from 2 to 1024. Each layer of the generator of StyleGAN accepts a 512-dimensional latent code as input. Due to the good semantic properties of the latent space of StyleGAN, many extensions on the latent space of StyleGAN have been born recently, such as $\mathcal{W}+$ and $\mathcal{S}$ space, and these researches are good to enhance the applications of StyleGAN. The $\mathcal{W}+$ space of StyleGAN consists of 18 512-dimensional latent codes, each corresponding to one of the layers of StyleGAN generator and serving as its input. The excellent performance of StyleGAN's $\mathcal{W}+$ space has driven advances in the field of GAN inversion. GAN inversion work [6, 9, 16, 31, 37, 39, 52] can well invert images into the $\mathcal{W}+$ space of StyleGAN, thus facilitating semantic editing of images. Contrastive Language-Image Pre-training (CLIP) [28] trains a large number of image-text pairs, providing a powerful image-text representation. By encoding the image and text into the space of CLIP, the similarity of the image text can be quantified.

3 THE TEXTCLIP FRAMEWORK

Based on the powerful image generation capability of StyleGAN [22] and the cross-modal text-image representation capability of CLIP [28], we propose TextCLIP, a unified approach for text-guided image generation and manipulation. We divide TextCLIP into three stages:

- **Stage 1.** Using a pretrained encoder, the image or random noise is mapped to the $\mathcal{W}+$ [1] space of StyleGAN model pretrained on the FFHQ dataset [21] to obtain an initial latent code w_0.
- **Stage 2.** The initial latent code w_0 is passed through the level-channel mapper to obtain the mapping latent code w_t.
- **Stage 3.** The mapping latent code w_t is then processed differently with the initial latent code w_0 depending on the task to obtain the style latent code w_s, which is the input of the generator of a pretrained StyleGAN to obtain the final image.

3.1 Overview

The global framework is shown in Figure 3. TextCLIP supports either random noise or image as input (random noise for text-guided image generation task and image for text-guided image manipulation task), and we use a pretrained encoder to map the input to the latent space $\mathcal{W}+$ of StyleGAN[1]. For image, we use e4e [37] as the pretrained encoder; for random noise, we use a pretrained mapping network in StyleGAN [22] as encoder. the process can be formulated as:

$$w_0 = E(g_0), \tag{1}$$

[1]In our experiments, we actually use StyleGAN2 [22].

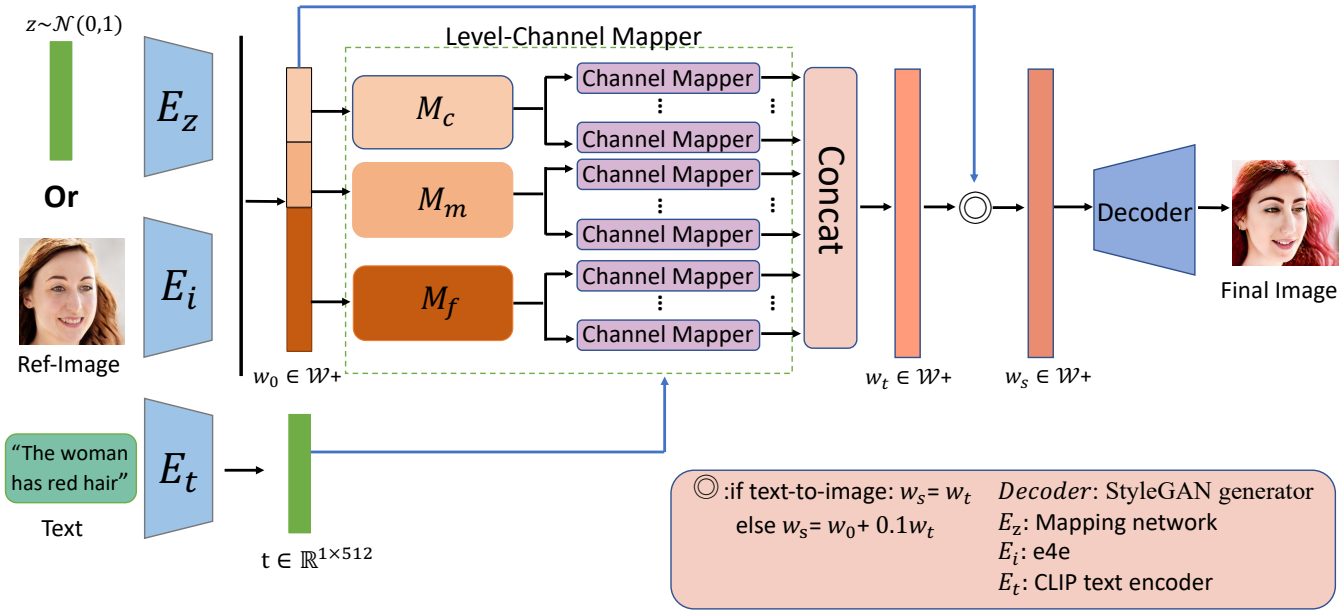


Figure 3: The framework of TextCLIP. The level mappers M_c, M_m, M_f consist of several fully connected layers that take a part of w_0 as input. There are a total of 18 channel mappers, each taking t encoded by the CLIP text encoder and the output of the corresponding level mapper as input. The outputs of the 18 channel mappers are concatenated to form w_t.w_t is then processed differently for different tasks to obtain w_s. w_s is used as input of the pretrained StyleGAN generator to obtain the final image.

where g_0 represents the initial input,w_0 represents the initial latent code mapped to the $\mathcal{W}+$ space of StyleGAN and E represents the pretrained encoder. The obtained initial latent code is then passed through the level-channel mapper to obtain the mapping latent code w_t, the mathematical equation of which is shown below:

$$w_t = F_{LCM}(w_0), \tag{2}$$

where F_{LCM} denotes the level-channel mapper. Next, we do different things depending on the task. For text-guided image generation task:

$$w_s = w_t, \tag{3}$$

For text-guided image manipulation task:

$$w_s = 0.1w_t + w_0, \tag{4}$$

where the style latent code w_s is used as the input of StyleGAN generator to obtain the final image g_s. The mathematical equation is shown below:

$$g_s = G(w_s), \tag{5}$$

where G denotes the generator of a pretrained StyleGAN, w_0, w_t and $w_s \in \mathcal{W}+$.

3.2 Level-Channel Mapper

The level-channel mapper consists of two parts: the level mapper and the channel mapper.

3.2.1 ***Level Mapper.*** Many previous studies have shown that different layers of StyleGAN generator control different attributes, so from coarse to fine we divided the layers of StyleGAN generator into three parts (coarse, medium, fine). In the same way we divided the input latent code w_0 into three parts, as follows:

$$w_0 = (w_0^c, w_0^m, w_0^f), \tag{6}$$

For each part, we design a network consisting of several fully connected layers, each of which is followed by operations such as layernorm and leaklyrelu. This is shown below:

$$M(w_0) = (M^c(w_0^c), M^m(w_0^m), M^f(w_0^f)). \tag{7}$$

In practice, we can train only one sub-network of M. Doing so allows us to change only the relevant image attributes and not some irrelevant ones.

As shown in Table 2, experimental results show that each layer of StyleGAN [22] controls different attributes, such as eye, hair color, age, face color, and other attributes. After our division, the coarse level controls attributes such as nose, head shape, lips, and hair length; the middle level controls attributes such as hair and face color; and the fine level controls age, gender and some micro attributes.

3.2.2 ***Channel Mapper.*** We design a channel mapper for each layer of a StyleGAN generator. There are 18 channel mappers in total. M^c corresponds to 4 channel mappers, $M^m$to 4 channel mappers and M^f to 10 channel mappers. For each channel mapper, it takes the output from the corresponding level mapper and the text code t encoded by the CLIP [28] text encoder as input. As shown in Figure 4, the text is first encoded by CLIP text encoder to obtain text conditional code t. t modulates the input that comes from the corresponding level mapper after processing in two fully connected

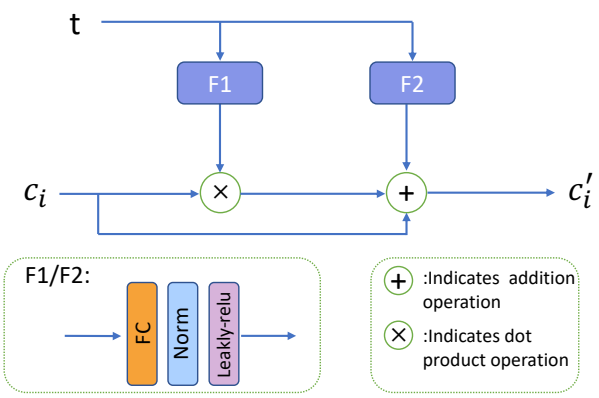


Figure 4: The structure of channel mapper.t is the text vector encoded by the CLIP text encoder. c_i is the input of the ith channel mapper and comes from the corresponding level mapper.

layers. The mathematical form is shown below:

$$c_i' = c_i + F_1(t)c_i + F_2(t), i = 0, 1, ..., 17, \tag{8}$$

where F_1 and F_2 are two networks designed by fully connected layers,c_i is the input of layer i. Finally, the resulting 18 channel styles are concatenated to obtain the final style latent code w_s. The mathematical form shown below:

$$w_s = Concat(c_0', c_1', c_2', ..., c_{17}'), \tag{9}$$

where $Concat$ means that the outputs of the 18 channel mappers are sequentially concatenated together.

3.3 Loss Function

3.3.1 ***Semantic Loss.*** An important aspect of text-guided image generation and manipulation is the need to ensure that the generated images are semantically consistent with the corresponding text. For this consideration, we propose semantic loss. The text and image are first encoded separately using CLIP [28] pretrained encoder, and then the result is computed as the cosine similarity to obtain the semantic loss.

$$\mathcal{L}_{semantic} = 1 - cos(t, F_{img}(G(w_s))), \tag{10}$$

where F_{img} represents the pretrained image encoder of CLIP, t is the text vector obtained by processing the CLIP text encoder, and cos represents the cosine similarity calculation, $\mathcal{L}_{semantic}$ represents semantic loss.

3.3.2 ***Identity Loss.*** We need to ensure that the generated image is identical to the original facial identity, so we introduce identity loss as follows:

$$\mathcal{L}_{ID} = 1 - cos(R(g_0), (R(G(w_s)))), \tag{11}$$

where g_0 represents the original image and R represents a pretrained Arcface [7] network for extracting the identity features of the image. The identity loss $\mathcal{L}_{ID}$ is obtained by calculating the cosine similarity of the face identity features of the two images.

Table 2: Layer-wise Analysis of a 18-layer StyleGAN Generator.

Level	Layers	Attributes
coarse	0-3	face shape,hair length,nose,lip,*et,al.*
medium	4-7	hair color,face color,*et,al.*
fine	7-17	age,gender,micro features,*et,al.*

3.3.3 ***Image Loss.*** The image loss consists of pixel loss $\mathcal{L}_{pixel}$ and image feature loss $\mathcal{L}_{lpips}$. Pixel loss refers to the fine-grained supervision of the generated image by comparing each pixel of the generated image with the original image g_0. Feature loss refers to the comparison of the images at the feature level, typically using a pretrained network for feature extraction [18]. Image loss is defined as follows:

$$\mathcal{L}_{pixel} = \|g_0 - G(w_s)\|_2^2, \tag{12}$$

$$\mathcal{L}_{lpips} = \|F_{VGG}(g_0) - F_{VGG}(G(w_s))\|_2^2, \tag{13}$$

where F_{VGG} represents a pretrained VGG network for extracting image features [18]. The total image loss is shown below:

$$\mathcal{L}_{img} = \lambda_{pixel}\mathcal{L}_{pixel} + \lambda_{lpips}\mathcal{L}_{lpips}, \tag{14}$$

where $\lambda_{pixel}, \lambda_{lpips}$ are the corresponding hyperparameters.

3.3.4 ***Fidelity Loss.*** After experimentation, it was found that previous research in text-guided image generation and manipulation tended to produce some low-quality and blurred images. To address this issue, we introduce the fidelity loss to prevent the generation of some low-quality and blurred images. It is shown as follows:

$$\mathcal{L}_d = \sigma(D(g_s))), \tag{15}$$

where σ represents sigmoid function, g_s represents generated image, D represents StyleGAN discriminator. We use a pretrained discriminator D of StyleGAN [22], which performs image fidelity determination to prevent the model from generating blurred photos.

3.3.5 ***Overall Loss.*** In summary, in order to make the images generated by the model realistic and semantically similar to the corresponding text, we define the following loss function:

$$\mathcal{L} = \lambda_{semantic}\mathcal{L}_{semantic} + \lambda_{ID}\mathcal{L}_{ID} + \lambda_{img}\mathcal{L}_{img} + \lambda_d\mathcal{L}_d, \tag{16}$$

where $\lambda_{semantic}, \lambda_{ID}, \lambda_{img}, \lambda_d$ are the corresponding hyperparameters.

4 EXPERIMENTS

4.1 Experiments Setup

4.1.1 ***Datasets.*** In order to carry out the performance of text-guided face image generation and manipulation, we conducted our experiments to verify the soundness and efficiency of the TextCLIP method. We have selected the following face dataset to carry out our experiments.

- **Multi-modal CelebA-HQ Dataset** [42]: a multimodal dataset consists of images, descriptive text, semantic masks and sketch,and contains 30,000 images, 24,000 images in the training set and 6,000 images in the test set. Each image of Multi-modal CelebA-HQ Dataset corresponds to 10 text descriptions.

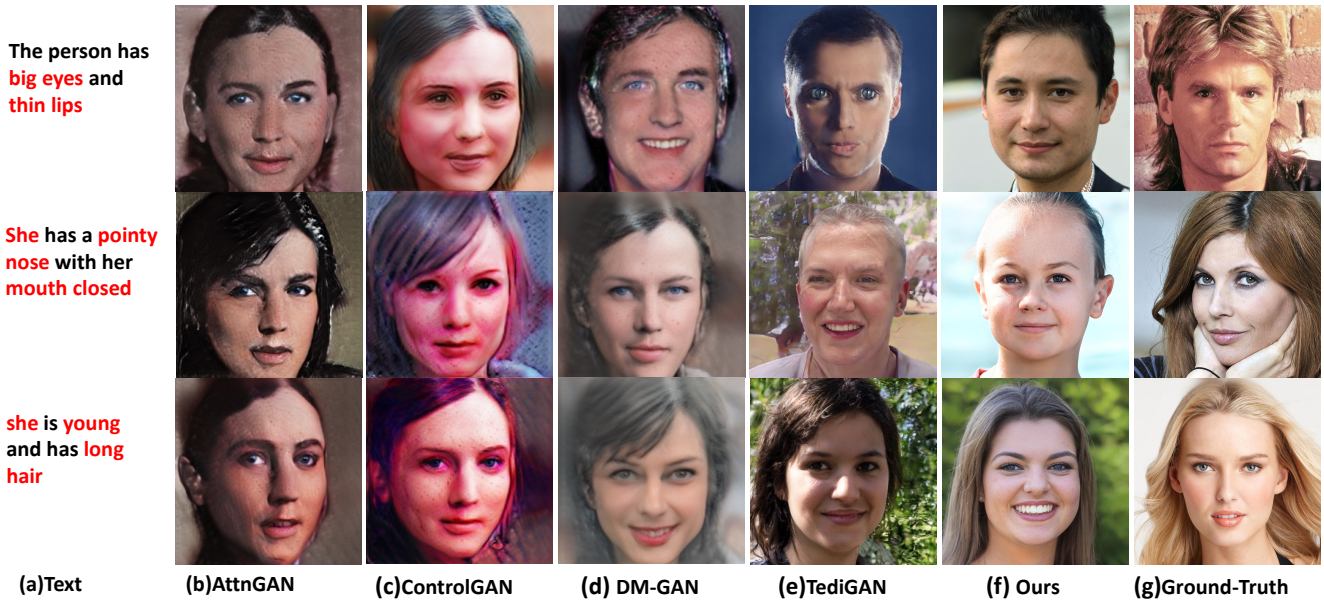


Figure 5: Qualitative comparison of text-gudied image generation compared with the state-of-the-art methods. TextCLIP generates more realistic and semantically similar images than previous methods.

4.1.2 ***Evaluation Metric.*** Text-guided image generation and manipulation require that the generated images are not only really enough to be realistic but also maintain a semantic similarity to the corresponding text. For this purpose, we have chosen the following evaluation metric.

- **Frechet Inception Distance (FID)** [13]: FID represents the distance between the feature vectors of the generated image and the feature vectors of the real image. The closer the distance is, the better the result of the model.FID gives us a good indication of whether the model is generating the exact data we desired.
- **R-Precision** [45]: another important property of text-guided image generation and manipulation is semantic similarity.we use R-Precision which evaluates the top-1 retrieval accuracy as the major evaluation metric in an image The higher the value of R-Precision, the higher the semantic similarity.
- **Learned Perceptual Image Patch Similarity(LPIPS)** [49]: to further evaluate the similarity of the generated image and the original image, we use LPIPS, which is a metric that learns the inverse of the generated image and the real image. A lower value of LPIPS indicates that the two images are more similar.
- **Identity similarity(IDS)** [15]: for text-guided image manipulation, we want the modified face image to be identity consistent with the original image, so we use IDS to evaluate this performance. IDS denotes identity similarity before and after editing calculated by Curricularface. The higher the IDS, the better the identity similarity.

User study: we also conducted a user study. 10 users from different backgrounds were selected and a user study was conducted by randomly generating 50 images under the same textual conditions.

Table 3: Quantitative Comparison of Text-Guided Image Generation on the Multi-modal CelebA-HQ dataset.

Method	FID↓	R-Precision↑	LPIPS↓
AttnGAN [45]	125.98	0.232	0.512
ControlGAN [23]	116.32	0.286	0.522
DFGAN [36]	137.60	0.343	0.581
DM-GAN [54]	131.05	0.313	0.544
TediGAN [42]	106.37	0.188	0.456
TextCLIP (ours)	**88.27**	**0.384**	**0.396**

User request to rank images generated by different models under the same conditions. The user study consisted of the following aspects:

- **Image realism**: to evaluate whether the generated images are realistic.
- **Semantic similarity**: for the image generation task, semantic similarity refers to whether the generated image is semantically consistent with the corresponding text; for the image manipulation task, semantic similarity refers to whether the model modifies the input image according to the specified text.

4.2 Results on Text-Guided Image Generation

4.2.1 ***Quantitative Results.*** As shown in Table 3, on the Multi-Modal CelebA-HQ Dataset [42], we compared the three metrics FID, LPIPS, R-precision with previous works. Based on the powerful image generation capability of StyleGAN [22] and the powerful image text representation capability of CLIP [28], our proposed TextCLIP surpasses the previous state-of-the-art approachs. Our

Figure 6: Qualitative comparison of text-guided image manipulation compared with the state-the-of-art methods. TextCLIP accomplishes more accurate semantic editing against the original image than previous methods.

Table 4: User Study on Multi-modal CelebA-HQ dataset. Acc. denotes semantic similarity and Real. denotes image realism.

Method	Acc. (%)↑	Real.(%)↑
AttnGAN [45]	18.6	12.8
ControlGAN [23]	19.7	13.9
DM-GAN [54]	21.1	16.3
TediGAN [42]	17.8	22.3
TextCLIP (ours)	**27.8**	**39.7**

Table 5: Quantitative Comparison and User Study of Text-Guided Image Manipulation on the Multi-modal CelebA-HQ Dataset. Acc. denotes semantic similarity and Real. denotes image realism.

Method	IDS↑	LPIPS ↓	Acc.(%)↑	Real.(%)↑
TediGAN [42]	0.18	0.45	10.8	12.4
StyleCLIP [26]	0.76	0.42	38.9	40.1
TextCLIP (ours)	**0.84**	**0.39**	**50.3**	**47.5**

proposed level-channel mapper can map textual information to the latent space $\mathcal{W}+$ of StyleGAN well and achieve high-quality image generation. At the same time, the loss function we designed can ensure generate the clearest possible images while ensuring semantic alignment. As shown in Table 4, user research shows that our approach outperforms the previous state-of-the-art approaches in terms of image realism and semantic similarity.

4.2.2 ***Qualitative Results.*** As shown in Figure 5, we compare qualitatively with the previous state-of-the-art methods. The comparison shows that our generated images have higher semantic similarity and image fidelity. In terms of semantic similarity, we use the semantic loss for supervision and exploit the powerful cross-modal text-image representation capability of the CLIP model to achieve higher cross-modal semantic alignment compared to other methods. In terms of image fidelity, we generated more realistic and realistic, higher resolution images. Unlike previous studies, we introduced an image fidelity loss to ensure that the generated images are realistic enough, taking into account the model's overfitting to semantic loss. Also based on the powerful generative power of StyleGAN, images with a resolution of 1024×1024 were generated. While AttnGAN [45] and ControlGAN [23] only can generate lower resolution images and TediGAN [42] sometimes generates some blurred images. Take the sentence "She has a pointy nose with her mouth closed" as an example, the focus is on "she", "pointy nose" and "mouth closed".Our generated images are highly semantically aligned with these three features; whereas TediGAN generated

images with mouths not closed, AttnGAN and ControlGAN generated somewhat blurred and low reslution images.As shown in Figure 2,for the same text, our method generates several different images, which demonstrates the diversity of our text-guided image generation methods.

4.3 Results on Text-Guided Image Manipulation

4.3.1 ***Quantitative Results.*** As shown in Table 5, we compared with the previous TediGAN [42], StyleCLIP [26]. Instead of using FID to evaluate text-guided image manipulation as in previous methods, we use IDS to evaluate whether the identity information is well preserved before and after the image is semantically modified,and use LPIPS to determine whether some semantically irrelevant image regions are preserved. And we conduct user study to determine the goodness of the model. The experiments show that, in contrast to previous methods, our proposed TextCLIP does a good job of semantically editing relevant image regions and partially preserving irrelevant image regions.

4.3.2 ***Qualitative Results.*** As shown in Figure 6, we compare it with the previous TediGAN [42], StyleCLIP [26]. Our method does a good job of modifying the semantically relevant parts according to the specified text, while not modifying the semantically irrelevant parts. In all six examples, TediGAN does not generate semantically relevant images well, while StyleCLIP produces similar results to our method, but the images produced by our method are more relevant to the given text while retaining the semantically irrelevant image regions well. This is not only because our designed level-channel mapper accurately maps the initial latent code according to the conditions of corresponding text, but also because our designed loss functions, including identity loss and semantic loss, accurately modify the images, preserving semantically irrelevant regions of the images such as face identity well.

5 ABLATION STUDY

5.1 Ablation Study On Loss Functions

As shown in Table 6, we designed a loss function that helps to improve the performance of the text-guided image generation and manipulation tasks. The semantic loss function makes the generated images semantically consistent with the given text, which takes advantage of CLIP [28] strong image-text representation capability. The identity loss function, especially on text-guided image manipulation tasks, allows for the good preservation of identity information of face images. The image loss function and the fidelity function allow the generated image to be close to the original image while being more realistic.

5.2 Ablation Study On Network Structures

As shown in Table 7, the level-channel mapper demonstrates a powerful performance combined with StyleGAN [22] and CLIP [28]. The level mapper helps to extract features in a hierarchical manner, and the channel mapper enables finer control of text-based conditions at a finer granularity. The experimental results show that the level-channel mapper formed by the combination of level mapper and channel mapper has excellent performance.

Table 6: Ablation Study On Loss Function. Gen. denotes image generation, Man. denotes image manipulation.

Method	Gen.		Man.	
	FID↓	R-precision↑	IDS↑	LPIPS↓
w/o $\mathcal{L}_{semantic}$	99.93	0.143	0.11	0.46
w/o $\mathcal{L}_{ID}$	90.34	0.433	0.34	0.44
w/o $\mathcal{L}_{img}$	94.54	0.428	0.78	0.45
w/o $\mathcal{L}_{d}$	93.28	0.483	0.83	0.40
TextCLIP (ours)	**88.27**	**0.384**	**0.84**	**0.39**

Table 7: Ablation Study On Network Structure. Gen. denotes image generation, Man. denotes image manipulation.

Method	Gen.		Man.	
	FID↓	R-precision↑	IDS↑	LPIPS↓
w/o Level Mapper	92.46	0.448	0.81	0.48
w/o Channel Mapper	100.22	0.396	0.78	0.42
TextCLIP (ours)	**88.27**	**0.384**	**0.84**	**0.39**

6 LIMITATIONS

After analysis we believe there are several limitations:

- TextCLIP is only done for specific face domains now, in the future we hope to extend this method to other domains such as flowers, birds, etc. In order to verify the superiority of the performance of our method on the flower and bird domains, we need to pre-train StyleGAN on the relevant flower and bird datasets. The StyleGAN pre-trained on the flower and bird dataset can generate high resultion flower and bird pictures, which is our next step in the future.
- Since TextCLIP is based on StyleGAN [22] and CLIP [28], the problems that arise in CLIP and StyleGAN itself will also arise in TextCLIP. For example, some attributes, such as hats and earrings, are not well represented in the latent space of StyleGAN so we do not get the desired results. In addition, CLIP is at risk of being attacked.

7 CONCLUSION

Based on the powerful image generation capabilities of StyleGAN and the image text alignment capabilities of Contrastive Language-Image Pre-training(CLIP), we propose a new approach that provides a unified framework for text-guided image generation and manipulation, does not require adversarial training, and can accept open-world texts . Extended experiments on the Multi-modal CelebA-HQ dataset demonstrate that our approach outperforms previous state-of-the-art methods in both text-guided image generation tasks and text-guided image manipulation tasks. In the future, we hope that TextCLIP will not be limited to the face domain, but will be extended to other domains such as flowers, birds, etc. In addition, for text-guided image manipulation tasks, we would like to explore a unified approach which does not need to go through the process of training different models for different classes of textual conditions, using only one model to complete the task.

REFERENCES

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4432–4441.

[2] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. 2021. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6711–6720.

[3] Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. 2021. Image-based clip-guided essence transfer. *arXiv preprint arXiv:2110.12427* (2021).

[4] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. 2020. RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10911–10920.

[5] Anton Cherepkov, Andrey Voynov, and Artem Babenko. 2021. Navigating the gan parameter space for semantic image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3671–3680.

[6] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. 2020. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5771–5780.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[8] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems* 34 (2021).

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2020. A disentangling invertible interpretation network for explaining latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9223–9232.

[10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544* (2021).

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[12] Yuchuan Gou, Qiancheng Wu, Minghao Li, Bo Gong, and Mei Han. 2020. SegAttnGAN: Text to image generation with segmentation attention. *arXiv preprint arXiv:2005.12444* (2020).

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[14] Xianxu Hou, Xiaokang Zhang, Yudong Li, and Linlin Shen. 2022. TextFace: Text-to-Style Mapping based Face Generation and Manipulation. *IEEE Transactions on Multimedia* (2022).

[15] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5901–5910.

[16] Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann. 2020. Transforming and projecting images into class-conditional generative networks. In *European Conference on Computer Vision*. Springer, 17–34.

[17] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. 2021. Talk-to-Edit: Fine-Grained Facial Editing via Dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13799–13808.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

[19] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems* 33 (2020), 12104–12114.

[20] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 34 (2021).

[21] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[23] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. 2019. Controllable text-to-image generation. *Advances in Neural Information Processing Systems* 32 (2019).

[24] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. 2020. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7880–7889.

[25] Daniil Pakhomov, Sanchit Hira, Narayani Wagle, Kemar E Green, and Nassir Navab. 2021. Segmentation in style: Unsupervised semantic image segmentation with stylegan and clip. *arXiv preprint arXiv:2107.12518* (2021).

[26] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.

[27] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1505–1514.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[30] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*. PMLR, 1060–1069.

[31] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2287–2296.

[32] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. 2021. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13960–13969.

[33] Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. 2021. StyleCLIPDraw: Coupling Content and Style in Text-to-Drawing Synthesis. *arXiv preprint arXiv:2111.03133* (2021).

[34] Mustafa Shukor, Xu Yao, Bharath Bhushan Damodaran, and Pierre Hellier. 2021. Semantic and Geometric Unfolding of StyleGAN Latent Space. *arXiv preprint arXiv:2107.04481* (2021).

[35] Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, and Zhenan Sun. 2021. Multi-caption Text-to-Face Synthesis: Dataset and Algorithm. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2290–2298.

[36] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. 2020. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865* (2020).

[37] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–14.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[39] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. 2020. Stylegan2 distillation for feed-forward image manipulation. In *European Conference on Computer Vision*. Springer, 170–186.

[40] Hao Wang, Guosheng Lin, Steven CH Hoi, and Chunyan Miao. 2021. Cycle-consistent inverse gan for text-to-image synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*. 630–638.

[41] Zongze Wu, Dani Lischinski, and Eli Shechtman. 2021. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12863–12872.

[42] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2256–2265.

[43] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Towards open-world text-guided face image generation and manipulation. *arXiv preprint arXiv:2104.08910* (2021).

[44] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2021. GAN inversion: A survey. *arXiv preprint arXiv:2101.05278* (2021).

[45] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1316–1324.

[46] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 833–842.

[47] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 5907–5915.

[48] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 1947–1962.

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[50] Xinjian Zhang, Yi Xu, Su Yang, Longwen Gao, and Huyang Sun. 2021. Dance Generation with Style Embedding: Learning and Transferring Latent Representations of Dance Styles. *arXiv preprint arXiv:2104.14802* (2021).

[51] Yutong Zhou. 2021. Generative adversarial network for text-to-face synthesis and manipulation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2940–2944.

[52] Bin Zhu and Chong-Wah Ngo. 2020. CookGAN: Causality based text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5519–5527.

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

[54] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5802–5810.